Studies in Correlative Assessing of Intrinsic and Extrinsic Indicators of Quality

Stefan Gradmann, Frank Havemann (both Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin), Jenny Sieber (Institut für Forschungsinformation und Qualitätssicherung)

Table of Content

Introduction

The assessment of research quality is one of the most important, yet one of the most difficult aspects of the scientific process. Evaluation procedures are in the center of many debates in academic, professional, and public policy circles. These debates are prevalent in the multidisciplinary field of education. The debates are due to the lack of consensus regarding the specific standards for assessing research quality and of a commonly agreed definition of the concept of quality in the field of educational research.

The traditional method of evaluation is the judgment by peers. Advantages and disadvantages have been extensively discussed in the literature[1]. The most often mentioned disadvantage of peer review is the problem that papers are judged on the reputation of the author instead of their quality.The process is time consuming and expensive and very often the review is performed either by narrow specialists who are actually unable to compare different projects, or by people with broad scientific qualifications, but without the specific insight required to evaluate the quality of a submitted paper. As a consequence, evaluation bodies tend to make use of quantitative methods they suppose to be more objective. The range of quantitative methods used in research assess is broad. The best known but most discussed methods came from the field of bibliometrics / scientometrics / webometrics. Indicators like the total number of articles published by an author, the h-index, g-index, the age-weighted citation ratio, impact factor and many more are used with the hope of reducing error and increasing accuracy of assessment.
Nevertheless, there are a lot of problems facing research quality assessment today. Two of them are the insufficient data coverage of Social Science and Humanities research publications in traditional bibliometric data bases on the one hand and the lack of a reasonable definition of the concept of research quality in the field of educational research on the other. The first makes the use of conventional bibliometric data sources highly disputable, the second has implications for the trust and fairness of peer judgements and the question of what actually should be measured. To overcome this problems a new approach based on the analysis of correlations between peer judgments and bibliometric measures was proposed and scrutinized in the EERQI project.

Related research activities

Citation analysis, as part of quality assessment tools, is limited by the bibliographic databases where citation data is gathered. This is the main target of criticism of the method. Citations in publications not indexed by these databases are simply lost. That is why new data sources need to be examined regarding their coverage and their usability for impact measures. Several researchers have investigated new quantitative methods for research impact evaluation to enhance traditional citation analysis. Among them Xuemei Li, Mike Thelwall and Dean Giustini in the end of 2011[2], Steve Kolowich[3], Jean-Claude Burgelmann and his colleagues[4], Jason Priem

[1] (Bornmann, 2008), (Bornmann, 2008), (Cicchetti, 1991), (Williamson, 2003).

[2] (Xuemei, Thelwall, & Giustini, 2011).

[3] (Kolowich, 2010)

[4] (Burgelman, Osimo, & Bogdanowicz, 2010).

and Bradely Hemminger - all in 2010[1], Mike Thelwall in 2003[2] and 2008[3] and Henk Moed in 2005[4]. Extracted from the literature, there are two principal directions: 1. the examination of WWW usage, and 2. citation analysis based on the WWW. The first named attempt evaluates the impact of a paper or a single researcher through potential readership statistics, e.g. article online views, clicks or downloads. The most prominent activity in this area is the project MESUR[5]. The project started in was funded by the Andrew W. Mellon Foundation and was conducted by the Digital Library Research and Prototyping team at Los Alamos National Laboratory. MESUR is not about one metric but a whole range of types and facets of usage metrics.

The second approach mentioned extends traditional citation analysis to the WWW. In an article published in 2001 Blaise Cronin argued that: "Citation analysis is an important piece of the bibliometric research pie; one that will become even more central with the growth of the web and for a very simple reason. The links (reference citations) provided routinely by authors in their reports and papers are a means of exposing the underlying socio-cognitive structure of science."[6] Making use of the infrastructure of the WWW, today's researchers have diverse options to communicate and disseminate their findings than ever before. These options include (open access) repositories, online journals, and Web 2.0 applications such as blogs, wikis, social bookmarking tools, Twitter and online reference management systems. Based on this infrastructure Cronin stated that: "After all, citations and 'sitations' are not merely similar phonetically ... Highly linked sites are the web's equivalent of highly cited papers."[7]

A third new trend occurred with the growth of reference management systems and their combination with social network features.[8] This third approach overlaps with web citation analysis, but intends to make use of the facilities that reference management systems can provide to track scholarly influence from users.

Regarding the problem of the lacking definition of research quality we think that the meaning and interpretation of research quality is strongly related with the intentions and purposes of the assessing body, as well as on the performance objectives, and the mission of the entity being evaluated. Determining the quality of

---

[1] (Priem & Hemminger, 2010).

[2] (Thelwall, 2003).

[3] (Thelwall, 2008)Thelwall, Mike, 'Bibliometrics to Webometrics', Journal of Information Science, 34 (2008), 605-621 <doi:10.1177/0165551507087238 >.

[4] (Moed, 2005).

[5] (Bollen, 2010).

[6] (Cronin, 2001), p. 2.

[7] (Cronin, 2001), p. 2.

[8] (Xuemei et al., 2011).

a piece of research necessitates scrutinizing the research processes and the research outputs. The most often used and best measurable research output is the dissemination of published research in the form of research articles.[1]

David Bridges, Professorial Fellow in the University of Cambridge Faculty of Education and Emeritus Fellow at St Edmund's College, Cambridge was member of the EERQI project team. He argued, that "quality assessment requires a judgement, a form of connoisseurship, based on a widely informed encounter with a situated text rather than anything which can be adequately captured by measurement".[2] Nevertheless this is exactly what the British Research Assess Exercise's (RAE) successor The Research Excellence Framework is aiming at in the future."It is widely expected that the ratings will initially be derived from bibliometric-based indicators rather than peer review. These indicators will need to be linked to other metrics on research funding and on research postgraduate training. In a final stage the various indices will need to be integrated into an algorithm that drives the allocation of funds to institutions."[3]

There have been several attempts to seize the relationship between academic impact measured via citations and research quality[4]. It was found that there exists a correlation between the assessment results of research output using bibliometrics and peer judgments. This is exactly what we were aiming to prove for the area of educational research.

Research Carried Out

Methodology
Intrinsic and extrinsic research quality indicators

Right from the beginning of the EERQI project it was clear that a stable and commonly agreed definition of the concept of research quality in the field of educational research was needed. The educational experts in the project agreed that the concept of research quality in educational research texts is rather difficult and complex, and for that reason it was decided to distinguish between intrinsic and extrinsic indicators of research quality of education research texts. What is integral to the quality of a text and what inherently constitutes elements of quality? What are the more indirect quality indicators of a research paper? The project team defined the terms as follows: Intrinsic indicators of the quality of a research text

[1] We did not take into account others forms of research output like oral contributions to a workshop,or lectures since the EERQI project proposal was aiming at the quality measurement of written research texts solely.

[2] (Bridges & Gogolin, 2011)

[3] (*The use of bibliometrics to measure research quality in UK higher education institutions*, 2007), p.2

[4] (Hornbostel, 1991), (Hornbostel, 2001), (Norris & Oppenheim, 2003), (Smith & Eysenck, 2002).

were those which were considered to be integral to the quality of that text, which are constitutive of that quality, which are a condition of judging it to be of high quality. Since quality consists e.g. in the coherence and consistency of the argument, and in the validity of the methods employed, the evidence of coherence, consistency or validity can be considered intrinsic indicators of the quality of the writing. Extrinsic indicators are those which do not inherently constitute elements of the quality of the piece, but which have a positive correlation with judgements based upon such elements. Extrinsic indicators correlate with the quality that can independently be discerned in the text. Extrinsic indicators have a "probabilistic" relation with quality.
The project retained originally had *rigour*, *originality*, *significance*, *integrity*, and *style* as intrinsic indicators of research quality. As a result of later discussions integrity and style were discarded as being too difficult to identify and only the first three indicators were actually retained. Mentions in online reference management systems, usage, and citation information were considered as relevant extrinsic quality indicators.

New data sources

Besides the traditional databases Web of Science and Scopus we suggested in 2010 the use of additional new data sources to calculate citation based metrics. The aim was to overcome the problem of lacking coverage of educational research published in other languages than English and in other formats than journal articles. Today, citations are no longer the only source of impact metrics and Web of Science is not longer the only database for bibliometric measures: the WWW itself can be mined for impact indicators. Jason Priem, researcher in the field of Information and Library Science and one of the first who investigated the viability of assessing scholarly impact over the social web instead of traditional citation analysis, stated in an article published in 2010: "Just as the early growth of Web–supported webometrics and usage–based metrics, the current emergence of "Web 2.0" presents a new window through which to view the impact of scholarship. These days, scholars who would not cite an article or add it to their Web pages may bookmark, tweet, or blog it. Arguably, these and related activities reflect impact and influence in ways that have until now eluded measurement."[1]

For the above reasons we proposed to work with online reference management tools. "Many scientists now manage the bulk of their bibliographic information electronically, thereby organizing their publications and citations material from digital libraries."[2] The use of online reference management systems like Mendeley, CiteULike, and Connotea is increasing continuously[3]. We think that these systems present an opportunity to create new data resources for quantitative measures. Metrics based on a diverse set of e.g. online reference management systems could yield broader, richer, and timelier assessments of current and potential scholarly impact.

---

[1] (Priem & Hemminger, 2010).

[2] (Hull, Pettifer, & Kell, 2008).

[3] (Priem & Hemminger, 2010).

Reference management software is a class of applications developed to assist in the process of compiling bibliographies and managing textual bibliographic records in one or more databases. Originally, beginning of the eighties, these applications were specifically conceived to facilitate the task of writing papers with all their bibliographic citations. Since a few years they have evolved significantly, and can be seen as a tool for the entire management of textual databases. Reference management systems like CiteULike, and Mendeley, have also incorporated social and collaborative features[1]. These features enable the users to share a personal library within a private or public group and to decide at what level to collaborate and be found by other researchers working in the same area. Users may also look for citations in the collective library that are similar to those stored in one's own library. By allowing researchers to expand their bibliographic records and eventually interact with other researchers in their field, collaborative reference management systems have a potential of growing into resource discovery environments.

In addition to the citation indicators based on Web 2.0 applications, Google Scholar and Web of science we aimed at the analysis of a second group of indicators: web usage. The advantages of usage data as part of impact measures lies in the chance to record interactions for all types of scholarly content, i.e. papers, journals, preprints, but also for blog postings, software etc. Since the measurement of these interactions can start immediately after the publication it is a very rapid indicator of scholarly trends[2]

Research design

In the course of the EERQI project a proposal for analysing the relation between assessment results based on extrinsic metrics and assessment results based on intrinsic indicators was made. The intrinsic indicators were operationalized and transferred into items of a peer review questionnaire. We intended to do a comparative and weighted analysis of a ranking based on the results of this scaled questionnaire and a ranking obtained from the extrinsic indicators in several iterations.

The underlying assumption was that there is a combination of extrinsic indicators, which best correlates with a combination of a combination of several or just one single intrinsic indicator. By discovering this combination of extrinsic indicators we were aiming at statements such as: The weighted combination of the extrinsic indicators "mentioning of article in Mendeley", "mentioning of article in Connotea", and "mentioning of article in citeUlike" corresponds best to the intrinsic indicator originality. Or: A ranking based on citations per paper gathered from Google Scholar weighted 2 times corresponds best to a ranking based on the average score on the indicator significance.

This part of the research strategy is illustrated in figure 1 below:

---

[1] (Duong, 2010).

[2] We are aware of the fact, that the need for rapid publication and citation of research information is more characteristic for the STM field than for e.g. the area of educational research.

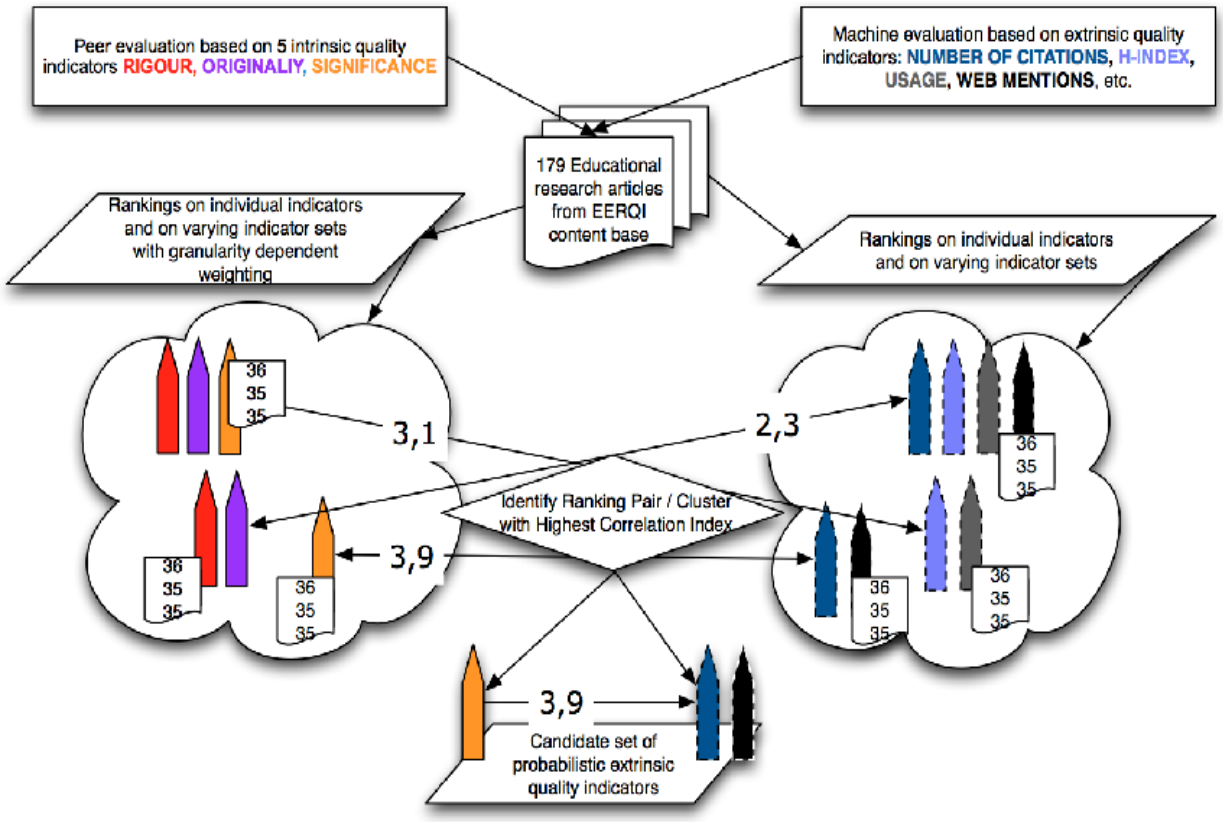


Figure 1: Correlation Identification Methodology Initial Steps

Furthermore, we had the intention to further process results from these initial steps in further iterations as depicted in figure 2 below:

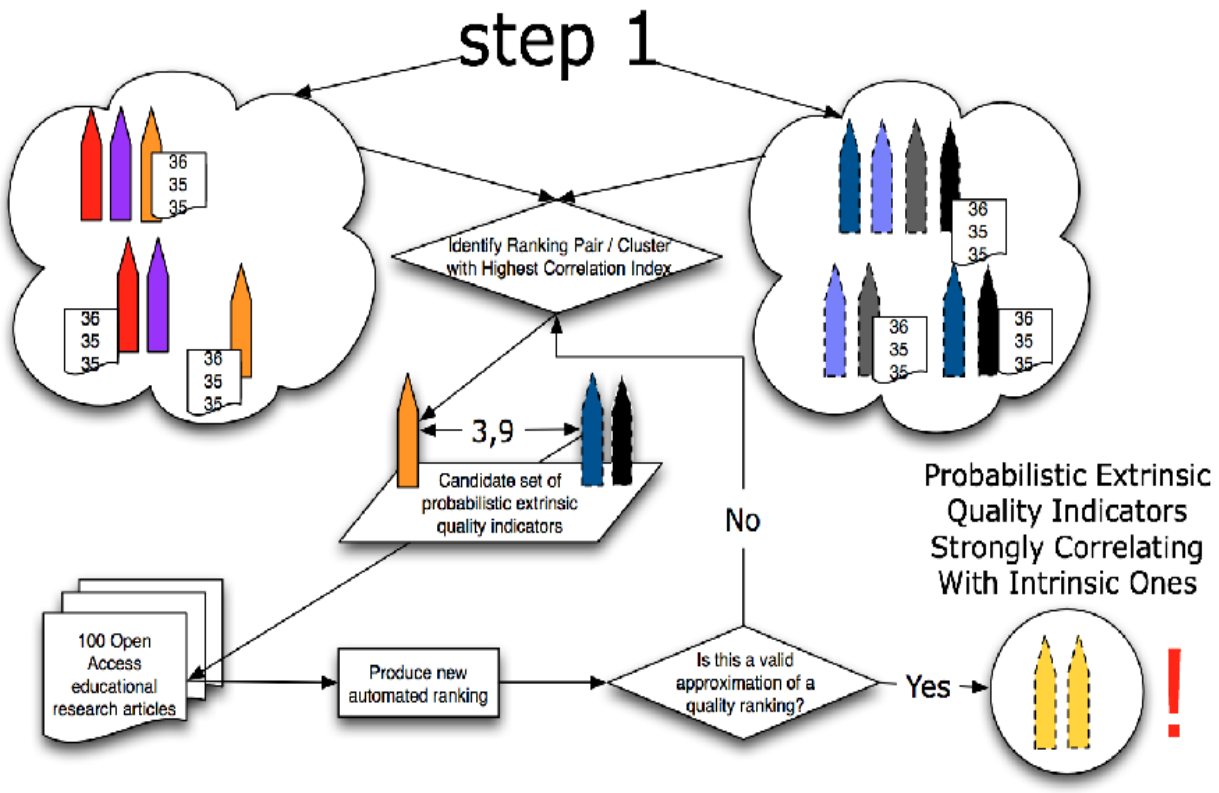


Figure 2: Further Iterations within the Correlation Identification Methodology

It should be noted that these further iterations were never carried out because of various synchronization problems within the project that resulted in important delays in the initial steps. The first results from these initial steps were not too inciting for further investigation, either (cf. below).

To obtain and compile the actual values for the above mentioned indicators we developed a piece of software called aMeasure. It is a stack of functions to measure extrinsic characteristics of research publications using Google Scholar, Google Web Search, MetaGer, LibraryThing, Connotea, Mendeley, and Citeulike[1]. In the context of the EERQI project aMeasure was used to collect information about extrinsic characteristics of educational research publications. It consists mainly of 4 parts:

1. a crawler to gather all information from Google Scholar (GS), Google Web Search and the Social Network Services,
2. a database to store the gathered information,
3. a client side application (JAVA-applet), and
4. a web interface to present the results and the content of the database to end users.

The main component of aMeasure is the crawler. For optimal work the crawler needs to be provided with author names. It has turned out that the major challenge in measuring extrinsic characteristics of research publications is the reliable identification of author names in the Social Network Services, GS, Google Web

[1] Cf. also (Stoye & Sieber, 2010).

Search, and MetaGer. We have therefore based our attempts on the findings presented by Derek Ruths and Faiyaz Al Zamal in the paper: "A Method for the Automated, Reliable Retrieval of Publication-Citation Records" published in 2010 . In this paper they present a series of filters to the results returned by an online publication search engine. One of these filters is a so-called name matching filter. Ruths and Zamal conducted several queries and retrieved "that when such a search is performed, the backend algorithm selects publications by applying a lenient filter to author names."[1] They found that slight modifications of the authors name have a significant impact on the initial set of candidate publications returned by the search engine and therefore recommended to use the following query syntax: author: "the first name of the author the initials of the middle names the last name of the author". Using this syntax the crawler queries GS for the authors and all of their papers. This is done via Screen-Scraping. In addition Google Web Search, MetaGer and the Social Network Services are queried to get information about the impact of each author's paper. The process of crawling is done on a central server located at Humboldt-Universität zu Berlin and it is constantly running in the background. As Google has limited the number of requests to an unknown randomly selected amount per IP per day the crawler is subject to this limit too. If this limit is reached and a user intends to search for an author's name which has not been already stored in the central database, a Java-applet is querying GS instead of the crawler.

All gathered data are stored in a central Mysql database located on the EERQI server to enable various exports via the web interface. GS is used to retrieve information about authors, their papers, and the citations of these papers. Due to the fact that Google does not provide an API aMeasure is required to use a technology called Screen-Scraping. The same technology is used to query MetaGer and the Social Network Services. A more comfortable method is used for retrieving results from Google Web Search and Mendeley, which are providing APIs to their search engines. These web search engines are queried with every single paper and the name of the author, for example: "Sahra Ahmed" + "Disablement following stroke". The results are then presented via a web interface.

Relying on the "name filter" solely is not a suitable, sufficient criterion to discern the publications that belong to a given author. Since many individuals share the same last name, many more share the same first name. Taking this into account we integrated a second filter, which ensures that the publications fall within the time span of an author's career. As we do not see how to get hold of each authors individual curriculum vitae we decided to limit the search results to the last 60 years arguing that an author is unlikely to start publishing before his/her 20th birthday and after his/her 80th year of life.

Besides we take into account the results of the so-called "classifier". The prototype of the classifier has been developed by ISN Oldenburg before the EERQI project and was refined and trained for Educational Science[2] content in the duration of the project. This classifier contains a fingerprint of those word shingles (strings of

[1] (Ruths & Al Zamal, 2010), p.3

[2] http://eerqi-classifier.projects.isn-oldenburg.de/

defined length), which are typical for professional and relevant publications in educational research. The classifier can be queried via an API for the probability of a given publication (identified by its URL) being from educational research or not.

We also considered the idea of making the results more precise via a matching of author names and affiliations or places. We decided not to take into account the affiliations as we see a problem of standardization of e.g. institutions names and change of institutions names in general in the data sources we are using. We also decided to abandon the plan to make use of author-place matching even if the problem of name standardization and name changing seems to be not that drastic according to e.g. names of cities. But since we need the full coverage of an author's publications for the calculation of e.g. the h-index the limitation of an author's publication to just one place of his career seems to result in a distorted picture. Taking into account the rapid movement of especially young researchers we would run the risk of losing a large amount of publications. Searches for e.g. "Stefan Gradmann" + "Berlin" resulted in much fewer hits than searching for "Stefan Gradmann" + "Hamburg", though we knew from the curriculum vitae that it is one and the same person in both cases.
The following extrinsic characteristics can be retrieved and calculated from GS using aMeasure:
• Number of papers per author.
• Number of citations per author.
• Year – first year of retrieved publication until last year of retrieved publication.
• Citations per year.
• Citations per paper.
• The h-index provides a single-number metric of an academic's impact. A scientist has index h if h of his/her Np papers have at least h citations each, and the other (Np−h) papers have at most h citations each. The h-index is calculated based on the full list of an authors output and the obtained citations. The h-index is robust in the sense that it is insensitive to a set of uncited or lowly cited papers but also it is insensitive to one or several outstandingly highly cited papers. This last aspect can be considered as a drawback and we therefore take into account the g-index.
• The g-index is an improvement of the h-index. It gives more weight to highly cited articles.[1]
• The e-index is aiming to differentiate between scientists with similar h-indices but different citation patterns.[2]
The following extrinsic characteristics can be retrieved and calculated from Google Web Search and MetaGer using aMeasure:
• Google Web Search hits matching the author's name.
• MetaGer hits matching the author's name.
The following extrinsic characteristics can be retrieved and calculated from Social Network Services using aMeasure:
• Citulike hits matching the author's name and the articles title.
• LibraryThing hits matching the author's name and the articles title.
• Connotea hits matching the author's name and the articles title.

[1] (Egghe, 2006).

[2] (Zhang, 2009).

• Mendeley hits matching the author's name and the articles title, readers of article in Mendeley.

Unfortunately, GS and Google Web Search present an estimated result count only, due to that every user and every API request is not able to see or get more than the first 1000 results for a specific search request. Regarding Google Web Search the company has shut down their old XML-API which enabled users to get very close to these 1000 results. Currently the Google-AJAX-API is limited to 64 search hits. If the Google Web Search reaches 64 hits, we are using "Screen Scraping" of Google Web Search to get the full list of results.

We learned that a robust method to identify authors is essentially needed as it is the critical step in making it possible to automatically track all the contributions that a researcher has made. This problem is very well known. In 2006 Elsevier launched its service "Scopus author identifier". The author identifier assigns a unique number to the authors who have published articles in journals covered by Scopus. An algorithm distinguishes those with similar or identical names on the basis of their affiliations, publication history, subject areas and co-authors[1]. Scopus excludes records from the process that lack sufficient data to determine a match. Once clearly identified, authors receive a unique identifier number. In 2007 CrossRef invited a number of people to discuss unique identifiers for researchers. In 2008 Thomson Reuters launched ResearcherID. ResearcherID tries to solve exactly the above illustrated problem. In the PLoS Comp Biol article Bourne and Fink argue that one solution to this difficulty is OpenID. OpenID is a standard. "That means that an identity can be hosted by a range of services and people can choose between them based on the service provided, personal philosophy, or any other reason. The central idea is that you have a single identity which you can use to sign on to a wide range of sites. There are two major problems with OpenID. The first is that it is poorly supported by big players such as Google and Yahoo. Google and Yahoo will let you use your account with them as an OpenID but they don't accept other OpenID providers. More importantly, people just don't seem to get OpenID" This state of our knowledge clearly isn't satisfactory and requires additional work in the future.

Currently aMeasure is filtering self citations with the help of GS. By using GS it is possible to search within all citations a paper has received. By subtracting all citations where the author of the original paper is also the author or co-author of the citing paper from the total amount of citations the paper has received we can filter out self citations. This technique prevents us from analyzing all citations manually, which would involve many queries to GS and would reduce the amount of papers and authors we are able to analyze per day. As some authors published a lot of papers which obtained many citations, and as there is a daily limit GS sets per user or IP per day this solution seems to be the most comfortable one in terms of returning hits in a reasonable time. From our point of view tools like CleanPoP do not seem to take this into account or present just a limited number of results concealing the illustrated problem of limited requests to Google. Besides, one further drawback of CleanPoP is the necessity to manually select author names and possible duplicates. This means that every single citing paper needs to be analyzed.

[1] (Qiu, 2008).

Source Data

The Publishing houses Symposium, VS-Verlag, Barbara Budrich Publishing, Taylor and Francis Publishing as well as the DIPF (German Institute for International Educational Research), IRDP (Institut de Recherche et de Documentation Pédagogique) and INRP (Institut National de Recherche Pédagogique) delivered nearly 6000 educational research publications (journal articles and book chapters) in the languages German, French and English and helped building the EERQI content base. Additional 42.000 educational research open access documents were crawled and added to the content base. Since most of the documents were in PDF format without sufficient metadata or XML-based structure, citation analysis within the EERQI content base could not be carried out as originally intended.

Analysis

For the analysis of correlation between intrinsic and extrinsic indicators a sample of 179 paper assessments based on intrinsic criteria was used in combination with two files of related extrinsic data:

- citation numbers of rated papers obtained with Google Scholar (on March 8, 2011)
- data from search engines and social-network services.

As the extrinsic author data generally suffered from homonymic authors we only used paper attributes. Papers were in English and in German and distributed over three thematic groups:

- Group 1 includes papers about "assessment, evaluation, testing & measurement" (35 / 35)
- group 2 about "comparative and inter-/multicultural education" (33 / 17)
- group 3 about "history and philosophy of education" (34 / 17)

We first had a closer look at the **interrelation between the three remaining intrinsic indicators** which each received a respective average of nine, three and four ratings of different aspects. This resulted in a combined rating score for each paper: the average ratings of all 16 aspects total score on a scale from 0 to 7.

The scatterplots in the three figures of mean scores of rigour, originality, and significance show that the latter two correlate best, especially for English-language papers. This is evident when comparing the low correlation strength in the interrelation of originality and rigour as shown in figure 3 below:

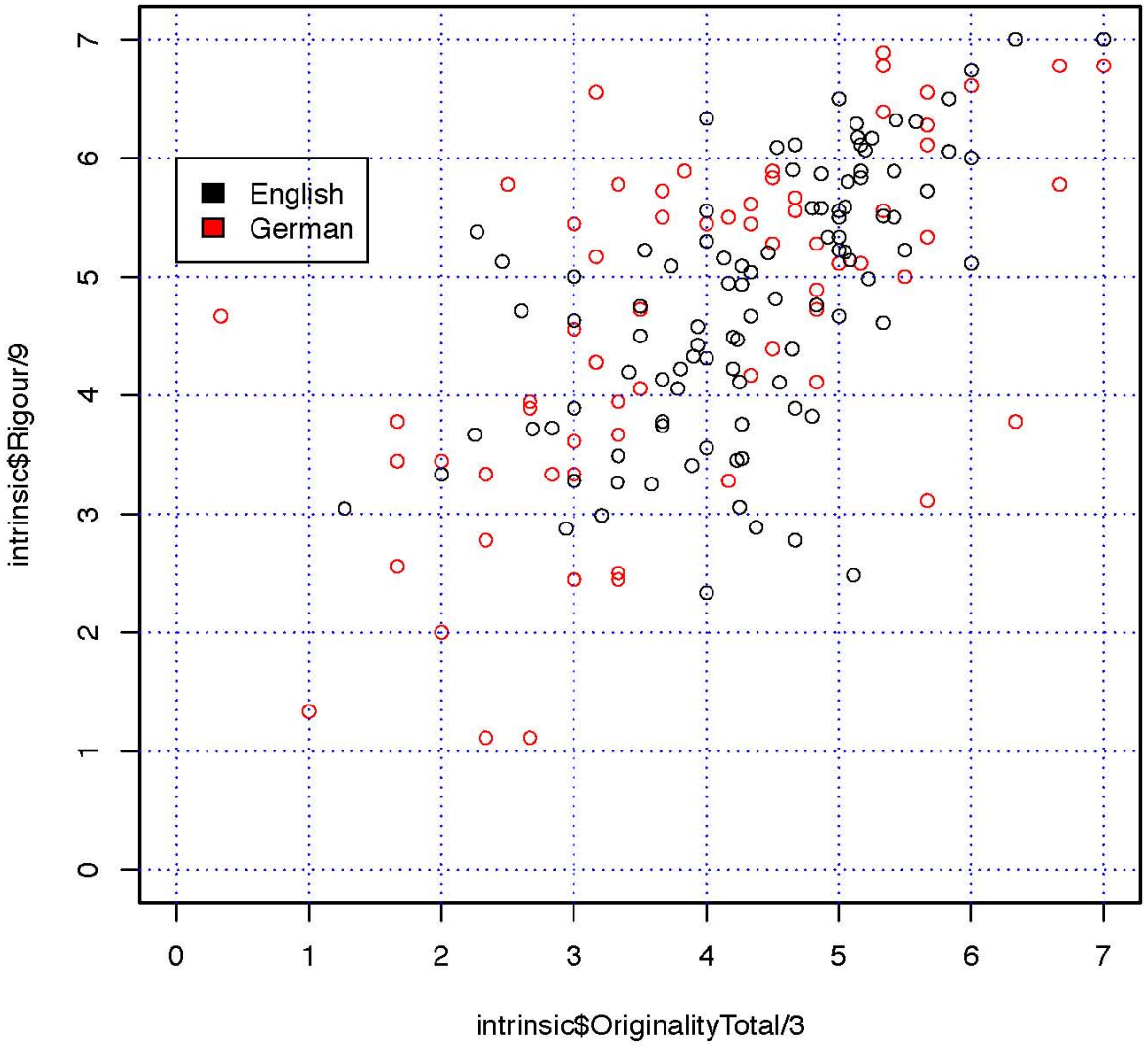


Figure 3: Originality - Rigour Interrelation

This clearly differs from the relatively high correlation strength in the interrelation of originality and significance as illustrated in the figure 4:

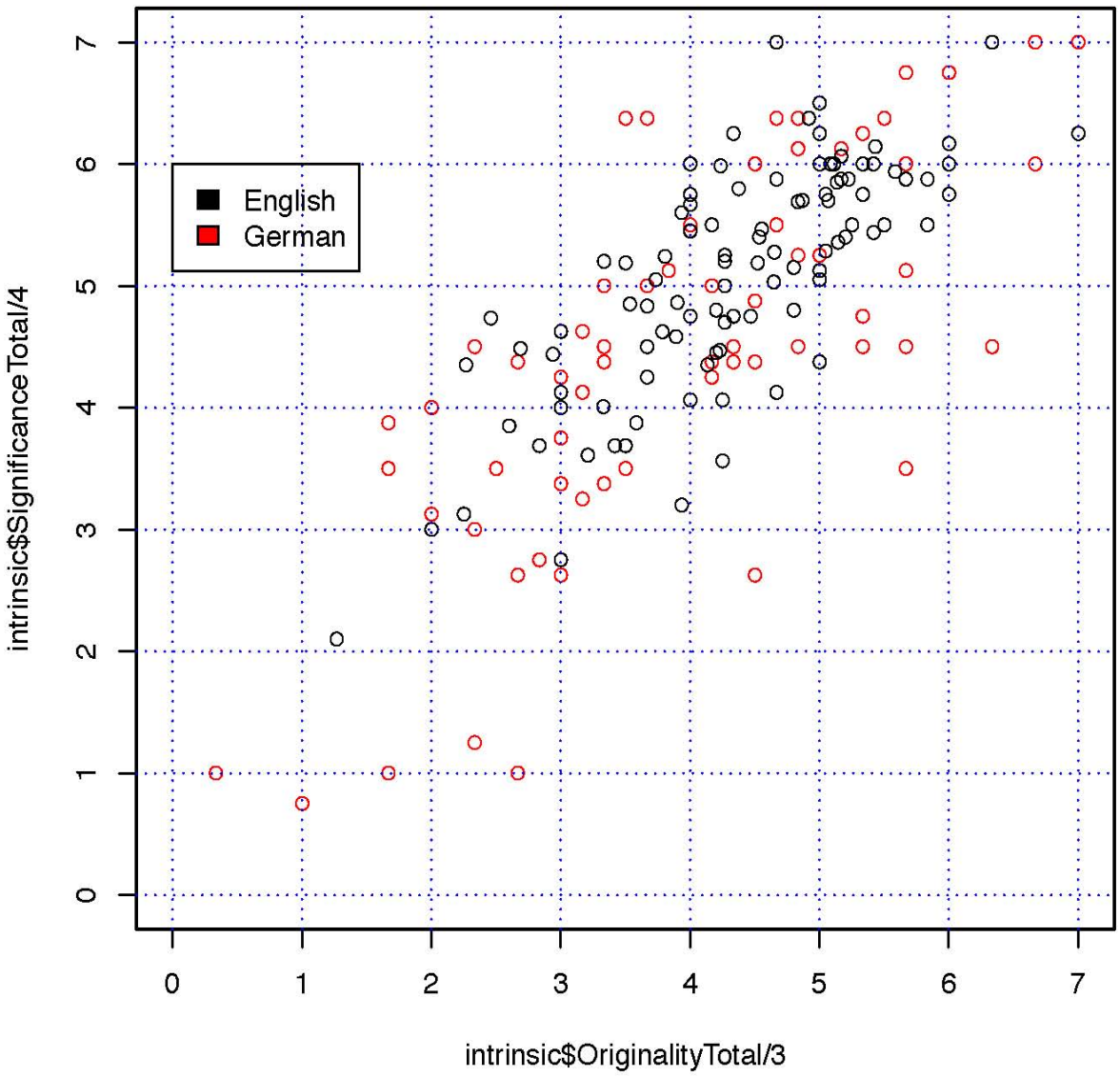


Figure 4: Originality - Significance Interrelation

Regarding the relative ratings in the three groups of papers it is interesting to note that the first group is clearly rated best as can be seen in figure 5 below (the values for the German papers do not differ significantly):

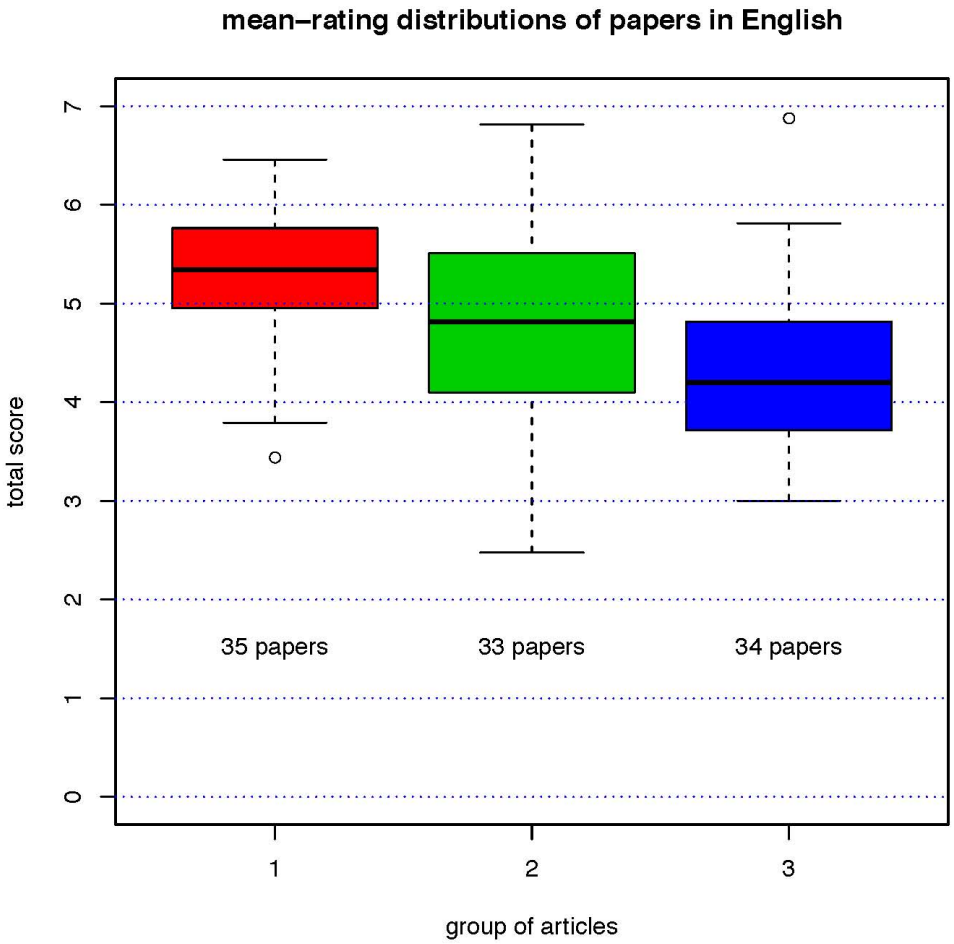
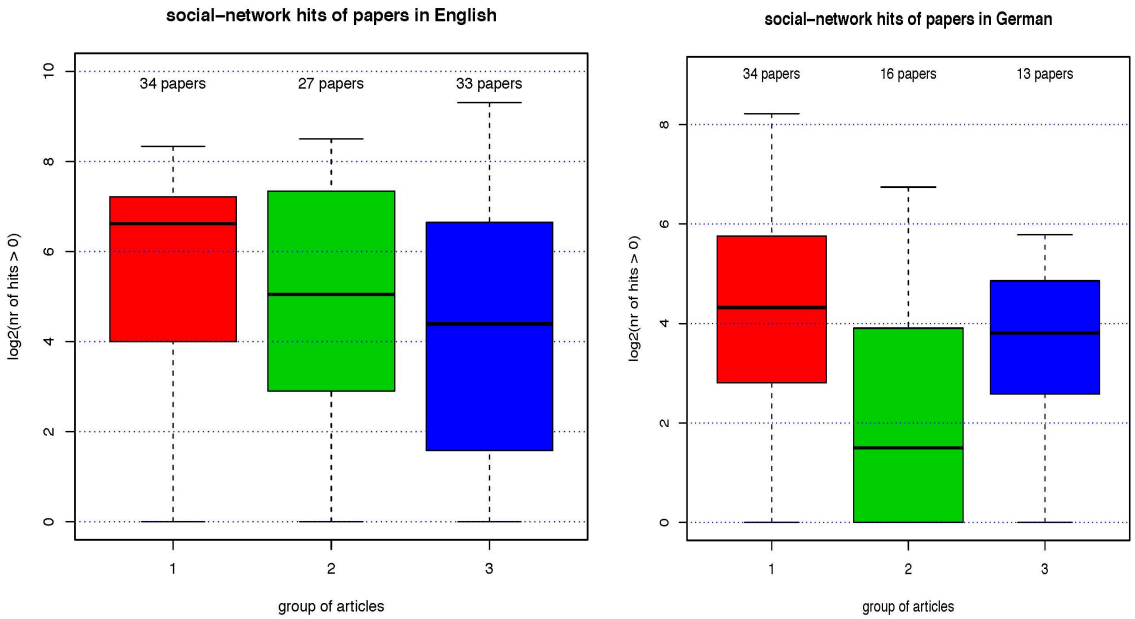
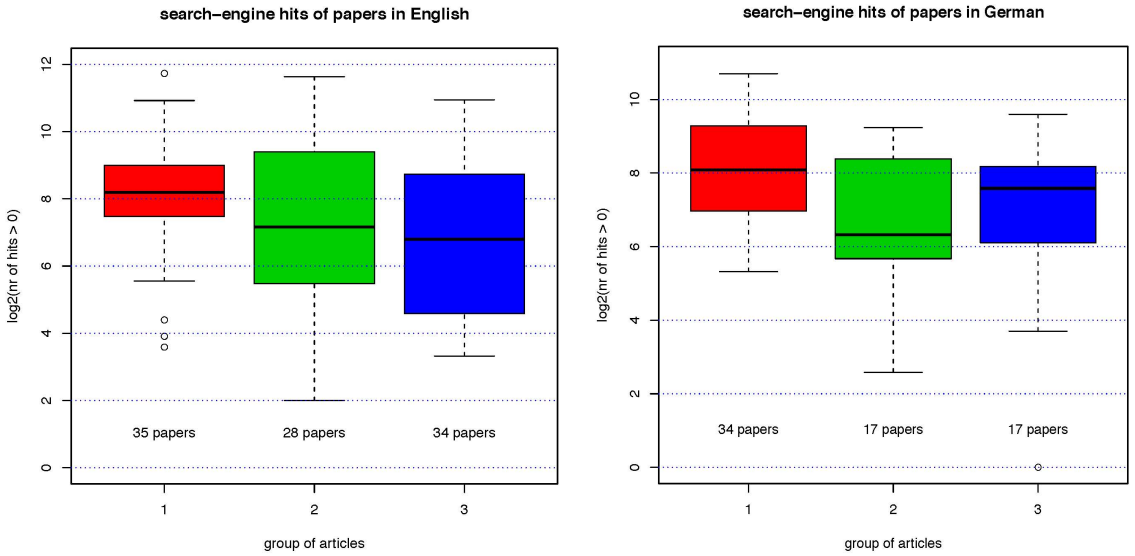


Figure 5: Boxplots of Mean Rating Distributions of Papers in English

We then looked into the **extrinsic paper data from search engines and social-network services**. These were extracted from the following sources: CiteULike, LibraryThing, MendReader, Google and Metager. Many papers had only hits in one service. To get useful data we therefore applied the in-dubio-pro-reo rule and selected maximum values. We also assumed that zero hits cannot be used as a valid value of an indicator and thus excluded papers without hits from the analysis. Furthermore, the hit distribution of papers with at least one hit was heavily skewed to the left: Many papers had only a few hits and only a few papers had many hits. We therefore used the logarithm of hit numbers as a more adequate representation. We use dual logarithms for all boxplot diagrams, i.e. the value of 8 on y-axis corresponds to 256 hits, a value of 10 to 1024 hits. The resulting diagrams show the following results:



Figures 6 and 7: Social Network Hits of Papers in English and German



Figures 8 and 9: Search Engine Hits of Papers in English and German

It can be observed here that all papers with social-network hits also have search engine hits and that both hit numbers correlate quite well in each of the three groups for papers in English - but and less well for papers in German.

Finally, we looked into **Citations in Google Scholar** and analysed the citation distributions for samples of the three groups. Not all papers were listed in Google

Scholar and only very few papers in German are in the sample: we decided to omit them. For the graphical representation we used the y-scale of dual logarithms of numbers of citation + 1. The addition of 1 is a usual bibliometric method to include papers without citations into the analysis of log-values. It can be justified with the argument that publishing a new result is its first citation.

The resulting figure 10 below shows the citation distributions for samples of the three groups:

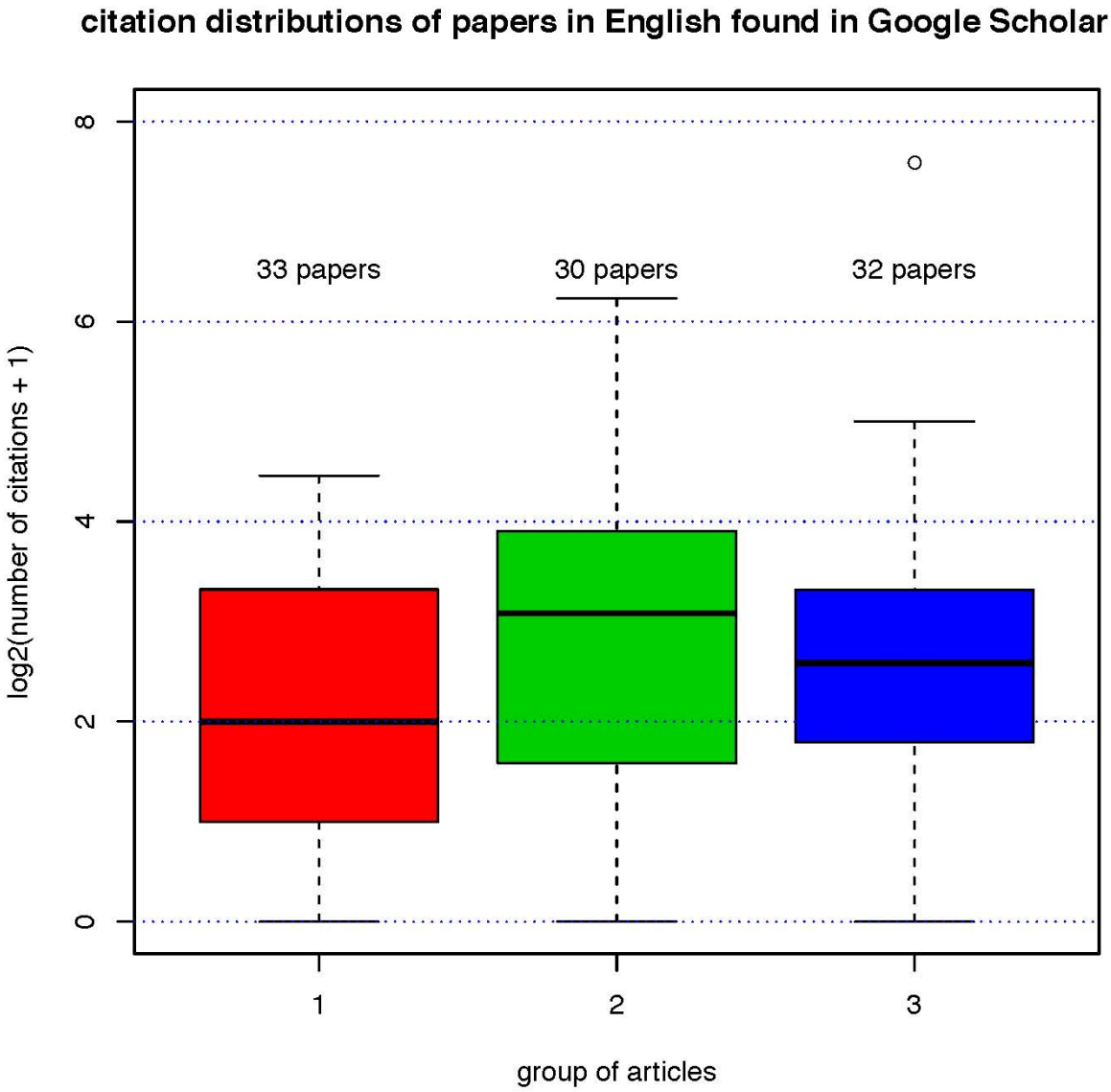


Figure 10: citation distributions for samples of the three groups

This diagram is interesting in that the first (red) group was rated best as could be seen in figure 5 but is cited worst (in contrast to the results for search engines and

social-network services, where for papers in English ratings and hit numbers on the aggregated level of thematic groups seem to correlate).

Results Assessment

Based on the selected articles we found no significant correlations between the extrinsic indicators of research quality and the intrinsic ones - we even found evidence of non-correlation! A first test based on a non-parametric regression model to analyse the correlation between the different indicators had not been successful, either. The measurement model with three intrinsic and two extrinsic latent factors which was conducted by Prof. Ton Mooij at Radboud University, revealed a significant inter-correlation between the extrinsic respectively the intrinsic group of indicators. The results give evidence that the indicators are multi-collinear. However, no significant correlations were found between the intrinsic and the extrinsic factors that were selected for this test. In a second attempt, rank correlations and conducting factor analysis calculations based on 179 articles were carried out. In the third approach, a test of modelling the correlation between the indicators by using different regression models (non- parametric) was not successful either. This first attempts to identify correlations between extrinsic and intrinsic indicators were primarily based on the testing of uni-variate and linear correlations between the two sets of indicators. Correlations between the multivariate elements of each set are most probably non-linear and complex.[1]

In any case, we can conclude that the two sets of indicators are not correlating significantly but that they rather are complementary to each other. In other words, an article that has been judged as of high quality referring the indicator "rigour" may be well presented in online reference management services, even if it was not considered to be 'original'. Extrinsic and intrinsic indicators as defined in the EERQI project can clearly complement, but not possibly replace each other.

A Look Ahead: Measuring monographs

Since we know, that scholarly monographs are not extinguish so soon[2] we are currently looking into ways to make monographs measurable. Since books are sufficiently covered neither by Web of Science[3] nor by Scopus[4], we decided to go for another group of tools - namely shared cataloging services like "Library Thing". But this will be reported on in a separate publication.

---

[1] (*EERQI Project Final Report*, 2011), p. 19

[2] (Wolfe Thompson, 2002).

[3] The Book Citation Index by Thomson Reuters was launched in the end of 2011 - after the official end of the EERQI project.

[4] Even if in 2011 325 book series were part of the Scopus database one can hardly mention this as sufficient coverage - even more if this is the number for the book series of all disciplines covered by the database. http://www.info.sciverse.com/scopus/scopus-in-detail/facts

Limitations

Our study is based on a small sample of documents. All of these were traditional journal articles.

Furthermore, even if the amount of data in the WWW allow us to get around the limitations of Web of Science' and Scopus' coverage, there is still the underlying problem of search engine reliability. Not only is there considerable variation between search engine retrieval performances, but the same search engine will also produce different results for the same search at different times and for different users. The situation is further aggravated by the fact that the coverage of Google is totally unknown up to know. That is why extreme caution is mandatory in using web-derived indicators for assessing research impact. Even more caution is advised when web based indicators shall be used in evaluation procedures.

The same, by the way, is true for the traditional sources of bibliometric information: the amount of fuzzy or simply wrong data and the lack of standardisation of attributes and their values we found in the course of our work is astonishing and makes us conclude that any figures derived from these sources needs to be used very cautiously, too, and any mechanistic trust in their reliability is likely to produce considerable harm!

Acknowledgements

Special thanks go to Prof. Ton Mooij for his contribution to the analysis of intrinsic and extrinsic data.

Literature

Bollen, J. (2010). The MESUR project: an overview and update. Retrieved January 20, 2012, Online: http://www.sparceurope.org/news/AAR_JB_MESUR_project_overview_update.pdf/view

Bornmann, L. (2008). Scientific peer review. An analysis of the peer review process from the perspective of sociology of science theories. *Human Architecture: Journal of the Sociology of Self-Knowledge*, *6*(2), 23-38.

Bridges, D., & Gogolin, I. (2011). The Process of Development of „Intrinsic Indicators". *EERQI Final Conference, Brussels, 15th–16thMarch 2011*.

Burgelman, J.-C., Osimo, D., & Bogdanowicz, M. (2010). Science 2.0 (change will happen… .). *First Monday*, *15*(7). Online: http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2961/2573

Cicchetti, D. (1991). The reliability of peer review for manuscript and grant submission. *Behavioral and Brain Sciences*, *1*(14), 119-186.

Cronin, B. (2001). Bibliometrics and beyond: some thoughts on web-based citation analysis. *Journal of Information Science*, *27*(1), 1-7. Online: http://jis.sagepub.com/content/27/1/1.full.pdf+html

Duong, K. (2010). Rolling out Zotero across campus as a part of a science librarian's outreach efforts. *Science & Technology Libraries*, *29*(4), 315-324.

EERQI Project Final Report. (2011). Hamburg. Online: http://eerqi.eu/sites/default/files/Final_Report.pdf#page=9

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131–152. Online: http://www.springerlink.com/content/4119257t25h0852w/fulltext.pdf

Hornbostel, S. (1991). "Drittmitteleinwerbungen. Ein Indikator für universitäre Forschungsleistungen?" *Beiträge zur Hochschulforschung*, (1), 57-84.

Hornbostel, S. (2001). "Third party funding of German universities. An indicator of re-search activity?" *Scientometrics*, *50*(3), 523-53.

Hull, D., Pettifer, S. R., & Kell, D. B. (2008). Defrosting the digital library: bibliographic tools for the next generation web. (J. McEntyre, Ed.)*PLoS computational biology*, *4*(10), e1000204. Public Library of Science. Online: http://dx.plos.org/10.1371/journal.pcbi.1000204

Kolowich, S. (2010). New Measures of Scholarly Impact. *Inside Higher Ed*. Online: http://www.insidehighered.com/news/2010/12/17/scholars_develop_new_metrics_for_journals_impact

Moed, H. (2005). The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science and Public Policy*, *34*(8), 575-583.

Norris, M., & Oppenheim, C. (2003). Citation counts and the Research Assessment Exercise V: Archaeology and the 2001 RAE. *Journal of Documentation*, *59*(6), 709-730.

Priem, J., & Hemminger, B. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, *15*(7). Online: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2874/2570

Qiu, J. (2008). Scientific publishing: Identity crisis. *Nature*, *451*, 766-767. Online: http://www.nature.com/news/2008/080213/full/451766a.html

Ruths, D., & Al Zamal, F. (2010). A Method for the Automated, Reliable Retrieval of Publication-Citation Records. *PLoS One*, *5*(8). Online: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0012133

Smith, A., & Eysenck, M. (2002). The correlation between RAE ratings and citation counts in psychology. Online: http://cogprints.org/2749/

Stoye, D., & Sieber, J. (2010). *Description of aMeasure: Measuring extrinsic quality indicators in educational research publications EERQI report* (pp. 1-8). Berlin.

Online: http://edoc.hu-berlin.de/oa/reports/reJ3Xv4PJ82ZM/PDF/29B6vgnyGba6.pdf

The use of bibliometrics to measure research quality in UK higher education institutions. (2007). London. Online: http://www.universitiesuk.ac.uk/Publications/Documents/bibliometrics.pdf

Thelwall, M. (2003). Web use and peer interconnectivity metrics for academic web sites. *Journal of Information Science*, *29*(1), 1-10.

Thelwall, M. (2008). Bibliometrics to webometrics. *Journal of Information Science*, *34*(4), 605-621.

Williamson, A. (2003). What will happen to peer review? *Learned Publishing*, *16*(1), 15-20. Online: http://www.ingentaselect.com/rpsv/cgi-bin/cgi?ini=xref&body=linker&reqdoi=10.1087/095315103320995041

Wolfe Thompson, J. (2002). The Death of the Scholarly Monograph in the Humanities? Citation Patterns in Literary Scholarship. *Libri*, *52*, 121–136. Online: http://www.librijournal.org/pdf/2002-3pp121-136.pdf

Xuemei, L., Thelwall, M., & Giustini, D. (2011). Validating online reference managers for scholarly impact measurement. *Scientometrics*, *89*(3), 1-11.

Zhang, C.-T. (2009). The e-Index, Complementing the h-Index for Excess Citations. *PLoS ONE*, *4*(5). Online: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.000